International Journal of Creative and Open Research in Engineering and Management
ISSN: 3108-1754 (Online)
Volume 02 Issue 11 November-2026 | Impact Factor: 3.5



BASIL: Bayesian Adaptive State Imagination and Learning with Live Telemetry Halting and In-Context Rule Induction

Saurabh Shinde[1], Prof. Vishal Joshi[2]

[1] Indian Institute of Science Education and Research Bhopal (IISERB), Bhopal Bypass Road, Bhauri, Bhopal 462 066, Madhya Pradesh, India

[2] Vishwabharati Academy's College of Engineering, Sarola Baddi, Jamkhed Road, Maharashtra 414201, India

Corresponding Author Email: shinde22@iiserb.ac.in

How to Cite this Article:

Shinde, S. (2026). BASIL: Bayesian Adaptive State Imagination and Learning with Live Telemetry Halting and In-Context Rule Induction. International Journal of Creative and Open Research in Engineering and Management, 02(11), 1-9. https://doi.org/10.55041/ijcope.v2i11.080

License:

This article is published under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

© The Author(s). Published by International Journal of Creative and Open Research in Engineering and Management.







https://doi.org/10.55041/ijcope.v2i11.080

ABSTRACT

Planning in unfamiliar environments requires an agent to infer governing rules, decide whether additional computation is useful, and determine when to stop searching. We study these questions in Multiverse, a procedural benchmark of two-player connection games with disjoint training, validation, and test rule families. A baseline pre-search value-of-computation predictor has zero-valued labels on 73% of sampled states and a held-out correlation of −0.004 with reference planning gain. BASIL combines three mechanisms: a rule-conditioned world model with boundary-aware convolutions, a discrete board-state carrier and a symbolic tactical horizon; Live Telemetry Halting (LTH), which predicts a completed-search action-value gap from intermediate search telemetry; and Bayesian Hypothesis Filtration (BHF), which maintains a posterior over finite rule hypotheses across episodes. In a three-seed validation study, full search scores 0.958 against a greedy opponent and 0.771 against a 100-rollout simulator-search opponent, compared with 0.549 and 0.424 for the fast policy. Against the latter opponent, LTH scores 0.646 using 8.2 rather than 33.9 model expansions per move. On the held-out test split, full-search scores are 0.938 and 0.722, respectively. In six-episode hidden-rule matches, posterior entropy falls from 6.49 to 1.69 bits and posterior-sampled search reaches a score of 0.757 in the final two episodes. These results support adaptive planning within the specified rule grammar, while comparisons with heuristic stopping rules and the cost of symbolic operations delimit the conclusions.

Keywords—Model-based reinforcement learning; Adaptive computation; Monte Carlo tree search; Optimal stopping; Bayesian rule induction; World models

I. INTRODUCTION

Model-based decision-making couples a representation of the current state with predictions about the consequences of candidate actions. AlphaZero, MuZero, and EfficientZero demonstrate the effectiveness of combining learned prediction with tree search [1–3]. Nevertheless, a fixed search budget allocates the same number of simulations to simple forced moves and difficult tactical decisions. Bounded rationality instead asks whether the expected benefit of a computation justifies its cost [4,5]. Expert iteration and adaptive imagination provide complementary ways to connect rapid policy inference with more expensive deliberation [6, 7].

Adaptive planning is particularly challenging when deployment rules differ from those encountered during training. The agent must preserve relevant state information through imagined transitions, transfer its policy across board geometries, and identify when additional search could change its action. If the rules themselves are hidden, it must also distinguish uncertainty about its learned dynamics from uncertainty



about the environment's mechanics.

We investigate these issues in Multiverse, a family of deterministic two-player connection games. Universes vary in board dimensions, gravity, boundary wrapping, blocked cells, and winning tetrominoes. Splitting by canonical rule family separates generalization across rules from interpolation across positions within the same game. The study includes a visible-rule setting and a hidden-rule setting in which evidence accumulates across several games in one universe.

The baseline reveals three obstacles. First, a pre-search value-of-computation (VOC) predictor trained from independent reference searches receives sparse and noisy labels. Second, shallow search can miss an opponent's immediate winning reply when an initially weak policy spreads visits over many actions. Third, dense coordinate-dependent heads, unsuitable boundary handling, and drift in latent occupancy impede transfer and reliable multi-step imagination.

BASIL addresses these obstacles through three components. The BASIL-3 world model pins rule information in its latent state, uses geometry-aware shared spatial processing, and combines learned representations with an exact discrete occupancy carrier and a rule-based tactical operator. Live Telemetry Halting (LTH) uses intermediate tree statistics to predict the action-value difference relative to the completed search. Bayesian Hypothesis Filtration (BHF) maintains the surviving members of a finite rule grammar and conditions imagination on posterior-sampled rules.

The contribution is the combination of these mechanisms and its evaluation on held-out rule families. We give a conditional bound for threshold-based stopping and establish the exactness of deterministic hypothesis filtering under an explicit observation model. The experiments compare full search, learned stopping, pre-search gates, simple anytime heuristics, and simulator-based reference agents. They show substantial gains over the fast policy, but also show that simple stopping heuristics remain competitive. The scope of the method is hybrid neural and symbolic planning within a known grammar; it does not establish rule discovery in an unrestricted environment.

II. LITERATURE REVIEW

Learned planning and expert iteration. AlphaZero uses policy and value learning to guide search in a known game simulator [1]. MuZero learns a representation and transition model for planning without requiring a complete observation-level reconstruction [2]. EfficientZero strengthens learning in limited-data settings through representation consistency and related training mechanisms [3]. Expert iteration connects search-based improvement with an amortized policy [6]. BASIL follows this broad structure while incorporating explicit board occupancy and finite rule hypotheses.

Adaptive computation. Rational metareasoning treats computation as a decision with a cost and an expected benefit [4, 5]. Learned metacontrol similarly adapts the amount of imagination to the decision [7]. LTH differs from a single pre-search gate by observing the search as it develops. Its supervision comes from checkpoints of the same completed trace. This removes disagreement caused by using separate reference searches, although the resulting target still inherits the errors and randomness of the completed search.

Uncertainty and rule inference. Randomized prior functions provide a mechanism for maintaining diversity in bootstrapped neural ensembles [8]. In BASIL, model uncertainty is supplemented by uncertainty over a finite set of rule hypotheses. BHF performs deterministic consistency filtering rather than learning a general-purpose rule discovery algorithm. Posterior sampling then exposes the planner to the remaining plausible rules.

III. METHODOLOGY

A. Benchmark: Rule-Family-Disjoint Procedural Games

A universe is a deterministic, two-player, zero-sum game on an $H \times W$ board, with $H, W \in \{5, 6, 7\}$, padded to 7×7. Its rules specify gravity $g \in \{0, 1\}$, horizontal and vertical boundary wrapping $w_h, w_v \in \{0, 1\}$, zero to three blocked cells B, and a winning shape set $S \subseteq \{\mathrm{I}, \mathrm{O}, \mathrm{T}, \mathrm{S}, \mathrm{L}\}$ containing one or two tetrominoes. Vertical wrapping is disabled when gravity is active. Under rotations and reflections, the five shapes yield 2, 1, 4, 4, and 8 orientations, respectively. A player wins by occupying a valid placement of any permitted shape. A full board without a winner is a draw. The game value is expressed from the perspective of the player to move.



Training, validation, and test proportions are 70%, 15%, and 15%. The split hashes the canonical family (H, W, g, w_h, w_v, S), identifying transposed gravity-free boards so that equivalent transposed families do not cross splits. Blocked-cell realizations are not a separate family key. A 30-game greedy balance filter rejects degenerate candidate universes. This protocol separates families within the specified grammar; it does not imply generalization to new rule primitives.

The observation $o_t \in \mathrm{R}^{14\times7\times7}$ uses planes 0–2 for the current player's stones, the opponent's stones, and empty cells. Planes 3–12 contain blocked cells, the active-board mask, gravity, two wrapping flags, and five shape flags. Plane 13 contains legal moves. In the hidden-rule setting, planes 5–12 are withheld, while board geometry, occupancy, and legal moves remain observable.

B. Failure Diagnosis: Noisy Pre-Search Computation Labels

The baseline first selects a fast-policy action a_{fast} and a planned action a_{plan}. Its VOC label is zero if they agree; otherwise it is the difference between independent reference estimates for the two actions. Under the simplifying assumption that each estimate has variance σ^2_{ref},

$$\mathrm{Var}\left[\hat{Q}^{(1)}_{\text{ref}}(a_{\text{plan}}) - \hat{Q}^{(2)}_{\text{ref}}(a_{\text{fast}})\right] = 2\sigma^2_{\text{ref}}. \qquad (1)$$

The reported baseline labels are zero on 73% of states, with small nonzero action-value differences of approximately 0.02. Estimation noise can therefore obscure the supervision signal. A gate using policy entropy and one-step ensemble probes achieves a held-out correlation of $r = -0.004$ with reference planning gain. This diagnosis concerns that baseline and target construction, rather than an impossibility result for pre-search prediction.

C. Tactical Cold Starts and Action Confounding

With approximately 26 legal moves, 48 simulations supply fewer than two root visits per action on average. A root blunder is exposed only if search reaches the opponent's winning reply. Under an idealized uniform choice among approximately 25 replies, the probability of selecting one particular reply is $1/25$. Actual PUCT selection is not uniform, so this calculation illustrates the exploration burden rather than giving a general PUCT failure probability. In the reported diagnostic, an untrained 48-simulation planner failed to block an immediate loss in 20 of 20 trials.

Coordinate-dependent dense heads further weaken transfer: flattening a board before a linear policy map does not enforce spatial weight sharing. Finally, terminal-action training can confound the existence of a threat with the value of the selected action. If winning states are observed only with their terminating winning action, a win head can exploit state-level threat features without learning which action completes the shape. Counterfactual action probes address this ambiguity.

D. Rule-Conditioned World Model

The agent comprises a representation $s_0 = h_\theta(o_0)$, a fast prediction head $(p, v, \ell) = f_\theta(s)$, and K bootstrapped dynamics members $(s'_k, w_k) = g_{\theta_k}(s, a)$. The ensemble uses fixed randomized priors [8]. Search operates on these representations and transitions, with tactical information supplied by the discrete rule-conditioned operators below.

1) Pinned invariants and spatial heads. Let $d = o_0[3{:}13] \in [0,1]^{10\times7\times7}$ contain the static rule planes and let $M = o_0[4{:}5]$ be the active-board mask. The final ten latent channels are copied from d:

$$s = \left[\mathrm{Scale}(z_{\text{dyn}}, M) \mid d\right] \in \mathrm{R}^{C\times7\times7}. \qquad (2)$$

Representation and dynamics networks preserve these channels at every imagined step. Latent consistency with a stop-gradient target is evaluated only on the $C - 10$ learned dynamic channels, preventing fixed rule planes from trivially reducing the consistency loss [3].

Ordinary circular padding wraps the full padded width, incorrectly connecting an active boundary to off-board padding on smaller boards. Boundary-aware processing instead identifies the edges of the active mask. For zero-filled one-cell shifts $S_{\rightarrow}$ and $S_{\rightarrow}$, the horizontal edge masks can be written as

$$E_R = (M - S_{\rightarrow}M)_+, \qquad E_L = (M - S_{\rightarrow}M)_+. \qquad (3)$$

Analogous vertical masks identify the top and bottom edges. For feature map x, the horizontal boundary contribution has the form

$$\Delta_h(x) = w_h \left(E_L \sum_c (x \odot E_R) + E_R \sum_c (x \odot E_L) \right), \qquad (4)$$

with broadcasting over columns; Δ_v is defined analogously over rows. These terms connect the actual active boundaries within residual processing. Masked



global maximum and mean pooling provide board context. Spatially shared policy logits use local features and a 1×1 output convolution, and legality excludes off-board and blocked cells. These choices remove coordinate-specific dense mappings; exact translation symmetry remains conditional on the boundary, gravity, and mask conventions of the game.

The action-conditioned win head pools features at the played coordinate using the one-hot action map A_a, together with global maximum and mean features. Whenever a training state has a winning move, an additional legal non-winning action on the same board is queried with target zero. This contrast holds the state fixed while changing the action, discouraging a state-only shortcut.

2) Discrete occupancy carrier. Let $b = (b_{\text{my}}, b_{\text{opp}}, b_{\text{empty}})$ denote binary occupancy. The first three dynamic channels, scaled to $z \in [0, 1)$, carry occupancy in a low-order binary component:

$$\text{Pack}(z, b) = \left(\frac{\lfloor 256z \rfloor}{256} + \frac{b}{512} \right) \odot M, \tag{5}$$

$$\text{Unpack}(s) = 2^{\left(256 s_{:3} - \lfloor 256 s_{:3} \rfloor \right)} \odot M. \tag{6}$$

For binary b, unpacking recovers the encoded bit exactly in exact arithmetic. The activation perturbation is less than 2^{-8}. The operation adds no channels and preserves parameter shapes, although this is not a guarantee that predictions are unchanged. Exact recovery in an implementation requires sufficient numerical precision and extraction before an intervening operation alters the packed values. The quantization itself is not an ordinary differentiable map.

For a legal placement, occupancy advances by exact placement and perspective exchange:

$$b' = \left(b_{\text{opp}}, \min(1, b_{\text{my}} + A_a), (b_{\text{empty}} - A_a)_+ \right). \tag{7}$$

The legal action map is recomputed from occupancy and rules, including column-fall constraints when gravity applies. The resulting virtual observation $o^{\wedge\prime} = [\, b' \mid d \mid \ell' \,]$ is encoded again. Dynamic features are coupled to the learned transition using

$$z'_{\text{manifold}} = 0.85\, h_\Theta(o^{\wedge\prime})_{\text{dyn}} + 0.15\, z'_k, \tag{8}$$

while static rule planes remain pinned. The reported correlation of step-one value predictions with ground-truth encodings rises from 0.404 to 0.973 with this coupling. The discrete carrier must be preserved through the transition and projection sequence; it is not an assertion that arbitrary latent processing preserves encoded bits.

3) Dihedral-torus tactical horizon. A bank $\mathbf{K} \in \{0, 1\}^{19 \times 1 \times 4 \times 4}$ contains the orientations of the five tetrominoes. Let P_{torus} be the rule-conditioned boundary-padding operator. Pattern responses and a legal threat map are computed as

$$C_{\text{pat}} = \text{conv2d}\left(P_{\text{torus}}(b', \ell'), \mathbf{K} \right), \tag{9}$$

$$H_{\text{tac}} = P_{\text{torus}}^{\top}\left(\text{conv_transpose2d}\left(M_{\text{threat}}(C_{\text{pat}}), \mathbf{K} \right) \right) \odot \ell'. \tag{10}$$

Here M_{threat} selects rule-permitted near-complete motifs, and the adjoint maps contributions through wrapped boundaries to board coordinates. The tactical diagnostic yields 100% precision and recall for immediate wins, forced blocks, and double-threat forks. This is an empirical diagnostic within the tested rule families rather than a universal correctness proof.

The occupancy transition, legality logic, and tactical bank explicitly encode mechanics of the benchmark grammar. Accordingly, "no simulator access" means no external environment-simulator queries during deployed search. It does not mean that the agent lacks an internal symbolic transition mechanism or prior knowledge of the rule grammar.

E. Live Telemetry Halting

1) Trace supervision and telemetry. For a maximum of T simulations, halting decisions occur at checkpoints $\mathcal{T} = \{0, k, 2k, \ldots, T\}$. With a fixed random stream, a search stopped at t is a prefix of its completed trace. Let a_t be the most-visited action at t, breaking ties by prior; at $t = 0$, use the fast policy. Let Q_T be the root action-value estimates after the full search. Each completed trace supplies labels

$$D_t = Q_T(a_T) - Q_T(a_t), \qquad A_t = \mathrm{I}[a_t = a_T]. \tag{11}$$

We use D_t for the reference action-value gap, called stopping regret in the reported experiments. Because a_T is selected by visit count rather than by maximizing Q_T, D_t need not be nonnegative. Its nonnegative counterpart is $R_t = (D_t)_+$. This distinction preserves the reported target while avoiding an unjustified sign assumption. Both targets are relative to the completed search, not to an optimal policy or true game values. A consistent convention for unvisited entries of Q_T is required when constructing labels.

At every checkpoint a 24-dimensional feature vector ψ_t summarizes prior statistics, elapsed budget, visit concentration, root value geometry, action stability,



and ensemble disagreement. Reported features include t/T, $\log(1+t)$, maximum prior probability, prior entropy, $|v_0|$, top visit share, the top-two visit gap, visit entropy, and whether the visit leader is unreachable within the remaining budget. Value features include $Q_t(a_t)$, a top-two Q gap, and root-value drift. Stability features track leader changes; uncertainty features summarize ensemble dispersion on the leading and expanded actions.

A two-headed predictor outputs $(\hat{D}_t, \hat{A}_t)$ and is trained with a smooth-L_1 loss and a binary cross-entropy loss, respectively. A threshold $\eta(\beta)$ is calibrated by replaying validation traces for a target compute fraction β. The regret head stops at the first checkpoint satisfying $\hat{D}_t \le \eta(\beta)$; the agreement variant uses its agreement prediction. Checkpoints at zero permit skipping search, while later checkpoints permit early termination. Collecting checkpoint labels requires no additional reference-search runs beyond the full trace, but telemetry and predictor evaluation still have computational costs.

2) A conditional threshold guarantee.

Proposition 1 (Completed-trace gap bound)**.** *Assume that* $\sup_{t \in \mathcal{T}} |\hat{D}_t - D_t| \le \delta$ *almost surely for some* $\delta \ge 0$. *For* $\varepsilon \ge 0$, *let* τ_ε *be the first checkpoint with* $\hat{D}_t \le \varepsilon$, *defaulting to* T. *Then*

$$\mathrm{E}\left[\left(Q_T(a_T) - Q_T(a_{\tau_\varepsilon})\right)_+\right] \le \varepsilon + \delta. \qquad (12)$$

For any checkpoint-valued comparator $\tau^\star$ *with* $D_{\tau^\star} \le \varepsilon - \delta$, *one also has* $\tau_\varepsilon \le \tau^\star$ *pointwise.*

Proof. If halting defaults to T, the action gap is zero. Otherwise $D_{\tau_\varepsilon} \le \hat{D}_{\tau_\varepsilon} + \delta \le \varepsilon + \delta$. Since the upper bound is nonnegative, taking a positive part and then expectation preserves it. At the comparator checkpoint, $\hat{D}_{\tau^\star} \le D_{\tau^\star} + \delta \le \varepsilon$, so the first threshold crossing cannot occur later. ■

This is a conditional threshold guarantee, not a proof that the learned policy solves a global optimal-stopping problem. The experiments do not establish the assumed uniform error bound. In particular, correlation alone does not certify it. When $a_t = a_T$, the same-trace label is identically zero. This eliminates the mismatch of two independent reference searches, but the full trace remains a noisy, potentially biased reference.

F. Bayesian Hypothesis Filtration

1) Filtering hidden rules. With known board geometry and blocked cells, the hidden mechanics range over $u = (g, w_h, w_v, S)$. There are six admissible gravity-and-wrap combinations and 15 one- or two-shape sets, giving at most 90 candidates. The same universe persists through a match of six games. At each observation, a candidate survives only if it agrees with the observed legal moves and terminal status. In ongoing games, a candidate predicting an already completed winning shape is rejected; at termination, the candidate must explain the observed winner and final placement. Draw observations must likewise be consistent with the candidate's terminal rule.

Let U_t be the surviving candidates and h_t the observed history. With a uniform prior and deterministic, complete consistency likelihood,

$$P_t(u) = P(u \mid h_t) = \frac{\mathrm{I}[u \in U_t]}{|U_t|}. \qquad (13)$$

This posterior conditions on observed transitions, legal masks, and outcomes, treating the action sequence as given. It does not use a probabilistic opponent policy to infer rules from the opponent's action preferences. Exact Bayesian conditioning requires all rule-dependent observations to be included consistently; partial checks yield a posterior only for the checked evidence.

Proposition 2 (Deterministic filtration)**.** *Suppose the true universe* $u^\star$ *belongs to* U_0, *observations are noiseless, and candidate checks exactly encode the specified observation model. Then* $u^\star \in U_t$ *for every* t, *the posterior is uniform on* U_t, *and*

$$H(P_{t+1}) = \log_2 |U_{t+1}| \le \log_2 |U_t| = H(P_t). \qquad (14)$$

Proof. The true universe assigns likelihood one to every observed event, so induction prevents its elimination. Each deterministic likelihood multiplies the uniform prior by a zero-or-one consistency indicator; normalizing therefore gives the uniform distribution on the survivors. Since $U_{t+1} \subseteq U_t$, entropy cannot increase. ■

2) Posterior-sampled imagination. For each ensemble member, a rule hypothesis $u_k \sim P_t$ conditions the root representation and its imagined transitions. At a fixed state-action pair, let Z be a model prediction and U the sampled rule. The law of total variance gives

$$\mathrm{Var}(Z) = \mathrm{E}_U\left[\mathrm{Var}(Z \mid U)\right] + \mathrm{Var}_U\left(\mathrm{E}[Z \mid U]\right). \qquad (15)$$



The second term describes variation between candidate rules and vanishes for a singleton posterior. The first term contains conditional model variation and is interpretable as epistemic uncertainty only to the extent that the ensemble is calibrated. Neither term guarantees that finite-search decisions are Bayes-optimal. Surviving candidates may also remain observationally indistinguishable under the actions actually played.

G. Experimental Protocol

We evaluate three independent training seeds, {0, 1, 2}, and paired evaluation on 12 held-out universes with both seats per seed. Validation families support method comparison and halting calibration. A separate test evaluation uses a one-shot, cryptographically preregistered protocol, holding test rule families apart from validation-based calibration.

The score is the win rate plus half the draw rate, with larger values preferred. The tables report 95% bootstrap intervals over universes and paired differences from the fast policy. Rounded differences are retained as reported and can differ by 0.001 from subtracting displayed scores. Universe-level pairing accounts for shared task difficulty; the resulting intervals characterize variation across the sampled universes rather than proving robustness throughout the rule grammar.

The fast policy performs no tree-search expansions. Full BASIL-3 search uses a nominal budget of 48 imagined simulations, with observed expansions per move lower because simulations and newly expanded nodes are distinct units. Opponents are a greedy rule-based player and `uct-100`, a true-simulator MCTS player with 100 rollouts. OracleSim uses true-rule simulator search and serves as a reference rather than a guaranteed upper bound at a finite budget. Adaptive comparisons include LTH-Regret, LTH-Agree, pre-search regret and VOC gates, confidence, disagreement, random gating, visit-share stopping, and an unreachable-leader rule.

Compute fractions of 10%, 25%, and 50% are calibration targets, not exact realized per-game fractions. Model expansions per move measure search effort; they exclude neural inference outside expansion, board re-encoding, tactical operations, hypothesis filtering, and training. Consequently, the tables support expansion-efficiency comparisons rather than claims about wall-clock speed or total FLOPs.

Pre-search entropy gate	–	9.0	0.1332	70.3%

Table I: Offline halting calibration on held-out validation rule families at a nominal 25% budget. "Gap" is the reported stopping-regret metric from (11). The pre-search regret correlation is measured at $t = 0$; LTH-Regret correlation uses live checkpoints.

Policy	r	Exp./move	Gap ↓	Agree. ↑
LTH-Regret	+0.564	8.9	0.0911	74.7%
LTH-Agree	–	9.0	0.0797	76.3%
Visit-share threshold	–	8.9	0.0911	73.3%
Unreachable-leader stop	–	9.1	0.0815	75.3%
Pre-search regret gate	+0.253	9.1	0.1288	70.3%
Fixed simulation budget	–	10.3	0.0815	75.3%

IV. RESULTS AND DISCUSSION

A. World-Model and Halting Diagnostics

Figure 1 shows training curves for the baseline and combined BASIL-3 architecture. The baseline fast-policy score against greedy is 0.083, compared with 0.549 for BASIL-3 on validation. This is a comparison of complete variants, not an isolated estimate of the benefit of any single architectural change.

The upgraded representation raises the pre-search VOC correlation from −0.004 to +0.610. For the LTH target, correlation increases from +0.253 at $t = 0$ to +0.564 with live telemetry; corresponding simulator-trace values are +0.070 and +0.438 (Fig-ure 2). These correlations involve different predic-tion settings and should not be interpreted as one controlled ranking of otherwise identical targets.

At approximately a 25% calibration budget (Table I), LTH-Regret yields a mean reported gap of 0.0911 at 8.9 expansions per move, compared with 0.1288 for the pre-search regret gate and 0.1332 for the entropy gate. LTH-Agree gives 0.0797 at 9.0 ex-pansions and 76.3% final-action agreement. Simple anytime rules are strong: visit-share stopping also gives 0.0911, and unreachable-leader stopping gives 0.0815. Thus live telemetry improves on the listed pre-search gates, but these aggregates do not estab-lish a universal advantage over heuristic stopping.

B. Online Play on Unseen Validation Families

Table II presents validation comparisons; Figure 3 plots the corresponding score-compute trade-offs. Against greedy, full BASIL-3 search scores 0.958 versus 0.549 for FastPolicy, with a reported paired gain of +0.410 and 95% interval [+0.208, +0.618].

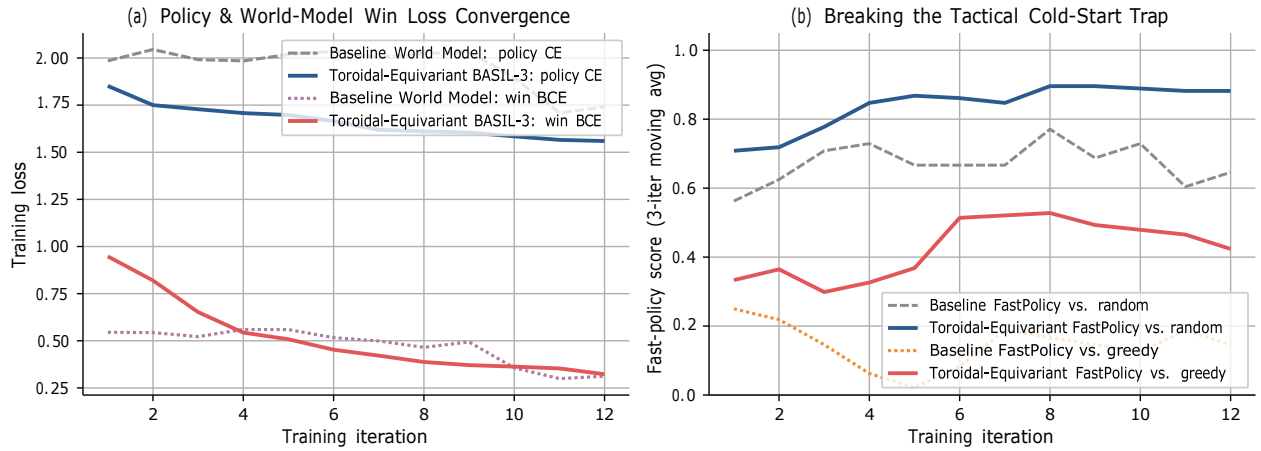


Figure 1: Training diagnostics. Left: policy cross-entropy and win binary cross-entropy for the baseline and BASIL-3. Right: fast-policy score against random and greedy opponents, shown with a three-iteration moving average. These curves compare the combined architecture changes and do not isolate each component's contribution.

OracleSim scores 0.924. The difference between the two search agents does not establish superiority over an exact simulator; search policies and finite budgets can produce different outcomes.

Against `uct-100`, full BASIL-3 scores 0.771 versus 0.424 for FastPolicy, a paired gain of +0.347 with interval [+0.194, +0.486]. LTH-Regret at the nominal 25% budget scores 0.646 using 8.2 expansions, versus 33.9 for full search: approximately 24% of its expansion count. This retains $0.222/0.347 \approx 64\%$ of the reported gain over FastPolicy. Confidence, disagreement, random, and pre-search VOC gates score 0.479, 0.493, 0.528, and 0.556, respectively.

The comparison with heuristic anytime rules is mixed. At a similar validation expansion count against `uct-100`, unreachable-leader stopping scores 0.653, slightly above LTH-Regret's 0.646. Against greedy, visit-share and unreachable-leader stopping score 0.854 and 0.861, compared with 0.743 for LTH-Regret at 25%. Wide intervals and the absence of paired intervals for every method-to-method comparison preclude a blanket significance claim.

C. *Held-Out Test Families*

Table III reports the separate test evaluation. Full BASIL-3 scores 0.938 against greedy and 0.722 against `uct-100`; FastPolicy scores 0.597 and 0.486. The paired gains are +0.340 with interval [+0.181, +0.500] and +0.236 with interval [+0.042, +0.403], respectively. LTH-Agree at 25% scores 0.681 against `uct-100`, with a paired gain of +0.194 and interval [+0.042, +0.347], at 10.5 expansions per move.

Heuristic stopping remains competitive on test: visit-share scores 0.688 at 8.9 expansions and unreachable-leader stopping scores 0.701 at 9.0 against `uct-100`. These values are comparable to, or above, the learned halting variants at similar budgets. The main supported conclusion is that adaptive search can improve substantially on the fast policy with fewer expansions than full search, while the best stopping rule depends on the evaluation setting.

D. *Hidden-Rule Adaptation Across Episodes*

Figure 4 and Table IV summarize six-game matches with hidden rule flags. Mean posterior entropy falls from 6.49 to 1.69 bits. The reported arithmetic mean candidate count falls from 90.0 to 4.6. These summaries are not interchangeable: exponentiating the mean final entropy gives approximately 3.2 effective hypotheses, whereas averaging the number of survivors gives 4.6.

Blind FastPolicy scores 0.565 over a match, and BHF plus FastPolicy scores 0.604. Full posterior-sampled search scores 0.715 overall and improves from 0.667 in Episodes 1–2 to 0.757 in Episodes 5–6. The adaptive variant scores 0.648 overall and rises from 0.556 to 0.674, using 19.6 rather than 38.3 expansions per move. The visible-rule reference scores 0.928 overall. BHF plus FastPolicy itself does not improve monotonically across the summarized episode groups, declining from 0.646 to 0.583. Thus shrinking rule uncertainty is guaranteed under the filtration assumptions, whereas monotonic per-

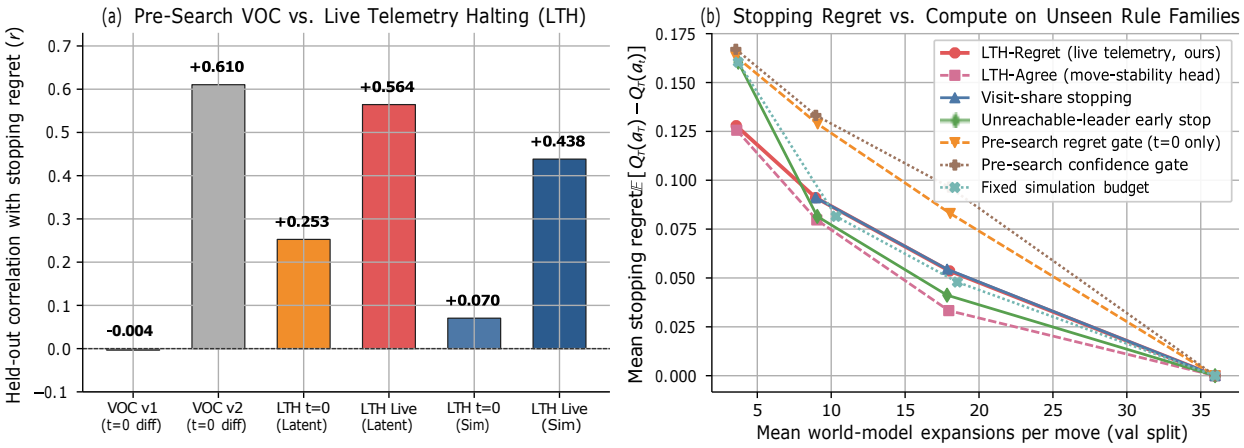


Figure 2: Halting diagnostics. Left: held-out correlations for the baseline and upgraded pre-search VOC predictors and the LTH predictors at zero and positive search depth. Targets and architectures differ across settings. Right: reported completed-trace action-value gap versus mean expansions on validation. The plotted quantity is a reference-search metric, not verified true-game regret.

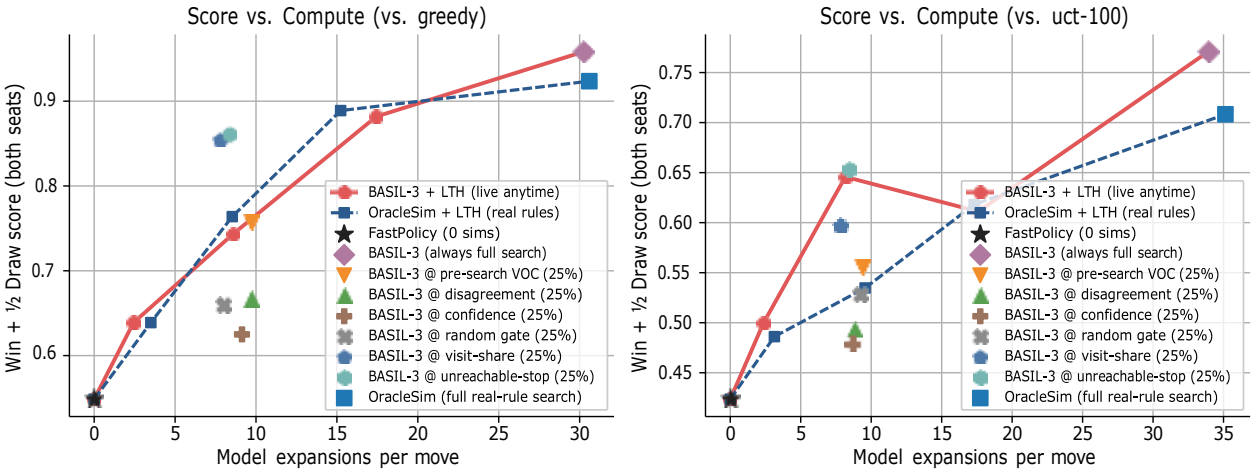


Figure 3: Reported validation score versus model expansions per move against greedy (left) and `uct-100` (right). Curves and markers compare learned halting, pre-search gates, heuristic stopping, full search, and OracleSim. Expansion counts omit the additional neural and symbolic overhead of each method.

formance improvement is not.

E. Discussion and Limitations

What the evidence supports. The strongest evidence is the consistent improvement of full BASIL-3 search over FastPolicy on validation and test families. Several adaptive policies retain part of this gain with fewer expansions. Live telemetry is informative for reference-trace prediction, and finite rule filtering reduces uncertainty across repeated exposure to the same universe. Together, these observations support the hybrid architecture as a test bed for adaptive planning under structured rule variation.

Learned versus encoded structure. The tactical horizon and occupancy update supply strong knowledge of connection-game mechanics. The method therefore combines learned latent prediction with symbolic operations. Its advantage over a less structured world model cannot be attributed solely to learned dynamics. Matched ablations are needed to isolate boundary handling, pinned rules, counterfactual probes, occupancy packing, re-encoding, and the tactical operator, including comparisons with a direct symbolic transition baseline using the same rule access.

Limits of the stopping reference. A completed 48-simulation search is neither an optimal agent nor an unbiased estimate of true action values. Agreement can preserve an incorrect decision, and the signed reference gap can be negative because final



Table II: Online play on unseen validation rule families. Scores, 95% bootstrap intervals over universes, paired differences from FastPolicy, and model expansions per move. Three seeds and both seats are included. Budget percentages denote calibration targets; realized compute varies.

Agent	Score ↑	95% CI	Δ vs. fast	Expansions/move
Opponent: `greedy`				
FastPolicy	0.549	[0.347, 0.750]	–	0.0
BASIL-3 (full search)	0.958	[0.875, 1.000]	+0.410	30.2
BASIL-3 + LTH-Regret (10%)	0.639	[0.438, 0.833]	+0.090	2.4
BASIL-3 + LTH-Regret (25%)	0.743	[0.597, 0.875]	+0.194	8.6
BASIL-3 + LTH-Regret (50%)	0.882	[0.757, 0.979]	+0.333	17.4
BASIL-3 + LTH-Agree (25%)	0.785	[0.632, 0.917]	+0.236	7.2
BASIL-3 + Pre-Regret (25%)	0.646	[0.465, 0.833]	+0.097	7.3
BASIL-3 + VisitShare (25%)	0.854	[0.715, 0.965]	+0.306	7.8
BASIL-3 + UnreachStop (25%)	0.861	[0.750, 0.958]	+0.312	8.4
BASIL-3 + Pre-VOC (25%)	0.757	[0.562, 0.938]	+0.208	9.7
BASIL-3 + Disagreement (25%)	0.667	[0.479, 0.847]	+0.118	9.7
BASIL-3 + Confidence (25%)	0.625	[0.403, 0.833]	+0.076	9.1
BASIL-3 + RandomGate (25%)	0.660	[0.458, 0.847]	+0.111	8.0
OracleSim (full search)	0.924	[0.826, 1.000]	+0.375	30.6
OracleSim + LTH-Regret (10%)	0.639	[0.458, 0.826]	+0.090	3.5
OracleSim + LTH-Regret (25%)	0.764	[0.618, 0.889]	+0.215	8.5
OracleSim + LTH-Regret (50%)	0.889	[0.778, 0.979]	+0.340	15.2
Opponent: `uct-100`				
FastPolicy	0.424	[0.208, 0.639]	–	0.0
BASIL-3 (full search)	0.771	[0.639, 0.889]	+0.347	33.9
BASIL-3 + LTH-Regret (10%)	0.500	[0.291, 0.722]	+0.076	2.4
BASIL-3 + LTH-Regret (25%)	0.646	[0.417, 0.854]	+0.222	8.2
BASIL-3 + LTH-Regret (50%)	0.611	[0.375, 0.812]	+0.188	17.6
BASIL-3 + LTH-Agree (25%)	0.583	[0.382, 0.778]	+0.160	7.9
BASIL-3 + Pre-Regret (25%)	0.542	[0.306, 0.764]	+0.118	8.1
BASIL-3 + VisitShare (25%)	0.597	[0.375, 0.806]	+0.174	7.8
BASIL-3 + UnreachStop (25%)	0.653	[0.465, 0.833]	+0.229	8.5
BASIL-3 + Pre-VOC (25%)	0.556	[0.354, 0.764]	+0.132	9.4
BASIL-3 + Disagreement (25%)	0.493	[0.292, 0.708]	+0.069	8.8
BASIL-3 + Confidence (25%)	0.479	[0.250, 0.701]	+0.056	8.7
BASIL-3 + RandomGate (25%)	0.528	[0.292, 0.750]	+0.104	9.2
OracleSim (full search)	0.708	[0.542, 0.861]	+0.285	35.1
OracleSim + LTH-Regret (10%)	0.486	[0.264, 0.708]	+0.062	3.1
OracleSim + LTH-Regret (25%)	0.535	[0.306, 0.750]	+0.111	9.6
OracleSim + LTH-Regret (50%)	0.618	[0.430, 0.806]	+0.194	17.3

action selection uses visits. The threshold proposition assumes uniform prediction error that has not been empirically certified. Strong heuristic stopping results also limit claims that a learned halting network is necessary. A more complete evaluation would compare calibrated positive-part gap, true-game outcomes, and uncertainty in paired differences at matched total cost.

Statistical and computational scope. Three seeds and 12 held-out universes per split provide a limited sample of the rule grammar. Bootstrap score intervals are often broad; method rankings should be interpreted accordingly. Mean expansion count omits representation, filtering, halting, and tactical costs. Wall-clock measurements, total FLOPs, memory, training cost, and hardware-matched comparisons are needed before claiming practical acceleration. Diagnostic claims of perfect tactical detection require counts, coverage, and reproducible test definitions.

Rule inference and reproducibility. Exact filtering depends on a finite correctly specified grammar, noiseless observations, and complete candidate checks. It does not address unknown primitives, stochastic transitions, or an excluded true rule. Rule distinguishability depends on the agent's ac-



Table III: One-shot evaluation on held-out test rule families. Scores, 95% bootstrap intervals over universes, paired differences from FastPolicy, and model expansions per move. Three seeds and both seats are included. Budget percentages denote calibration targets; realized compute varies.

Agent	Score ↑	95% CI	Δ vs. fast	Expansions/move
Opponent: `greedy`				
FastPolicy	0.597	[0.417, 0.764]	–	0.0
BASIL-3 (full search)	0.938	[0.833, 1.000]	+0.340	34.1
BASIL-3 + LTH-Regret (10%)	0.646	[0.451, 0.833]	+0.049	3.8
BASIL-3 + LTH-Regret (25%)	0.750	[0.597, 0.889]	+0.153	10.8
BASIL-3 + LTH-Regret (50%)	0.778	[0.625, 0.917]	+0.181	20.1
BASIL-3 + LTH-Agree (25%)	0.764	[0.597, 0.910]	+0.167	10.9
BASIL-3 + Pre-Regret (25%)	0.681	[0.500, 0.847]	+0.083	9.2
BASIL-3 + VisitShare (25%)	0.806	[0.639, 0.931]	+0.208	9.8
BASIL-3 + UnreachStop (25%)	0.743	[0.548, 0.917]	+0.146	9.1
BASIL-3 + Pre-VOC (25%)	0.743	[0.583, 0.896]	+0.146	10.7
BASIL-3 + Disagreement (25%)	0.604	[0.403, 0.792]	+0.007	7.2
BASIL-3 + Confidence (25%)	0.688	[0.542, 0.833]	+0.090	13.8
BASIL-3 + RandomGate (25%)	0.639	[0.417, 0.847]	+0.042	8.2
OracleSim (full search)	0.903	[0.784, 1.000]	+0.306	33.7
OracleSim + LTH-Regret (10%)	0.604	[0.403, 0.792]	+0.007	4.0
OracleSim + LTH-Regret (25%)	0.771	[0.583, 0.931]	+0.174	9.6
OracleSim + LTH-Regret (50%)	0.847	[0.708, 0.958]	+0.250	18.1
Opponent: `uct-100`				
FastPolicy	0.486	[0.292, 0.667]	–	0.0
BASIL-3 (full search)	0.722	[0.542, 0.875]	+0.236	35.8
BASIL-3 + LTH-Regret (10%)	0.542	[0.375, 0.708]	+0.056	3.7
BASIL-3 + LTH-Regret (25%)	0.639	[0.458, 0.819]	+0.153	10.0
BASIL-3 + LTH-Regret (50%)	0.688	[0.500, 0.854]	+0.201	18.8
BASIL-3 + LTH-Agree (25%)	0.681	[0.521, 0.833]	+0.194	10.5
BASIL-3 + Pre-Regret (25%)	0.556	[0.361, 0.736]	+0.069	8.4
BASIL-3 + VisitShare (25%)	0.688	[0.458, 0.861]	+0.201	8.9
BASIL-3 + UnreachStop (25%)	0.701	[0.542, 0.854]	+0.215	9.0
BASIL-3 + Pre-VOC (25%)	0.632	[0.403, 0.847]	+0.146	9.9
BASIL-3 + Disagreement (25%)	0.528	[0.333, 0.708]	+0.042	7.7
BASIL-3 + Confidence (25%)	0.507	[0.277, 0.736]	+0.021	11.5
BASIL-3 + RandomGate (25%)	0.549	[0.375, 0.722]	+0.062	6.9
OracleSim (full search)	0.708	[0.542, 0.847]	+0.222	36.0
OracleSim + LTH-Regret (10%)	0.514	[0.333, 0.681]	+0.028	3.2
OracleSim + LTH-Regret (25%)	0.653	[0.444, 0.833]	+0.167	9.7
OracleSim + LTH-Regret (50%)	0.708	[0.535, 0.868]	+0.222	18.1

tions, and posterior sampling alone does not guarantee information-seeking exploration. The present evaluation is summarized by aggregate outcomes. Releasing full configurations, raw seed-level match outcomes, bootstrap details, and the preregistration record would strengthen independent reproducibility and allow more precise paired comparisons.

V. CONCLUSION

BASIL combines a structured world model, search-telemetry halting, and finite Bayesian rule filtering for planning across procedural connection games. The reported results show substantial full-search gains over a fast policy and useful trade-offs between score and model expansions. Hidden-rule matches demonstrate uncertainty reduction and improved late-match performance for posterior-sampled search. The method's claims are bounded by its explicit rule grammar, symbolic tactical knowledge, completed-search supervision, competitive heuristic stopping baselines, and limited cost and uncertainty reporting. These boundaries motivate evaluation on larger rule spaces, stronger matched baselines, and complete end-to-end compute accounting.

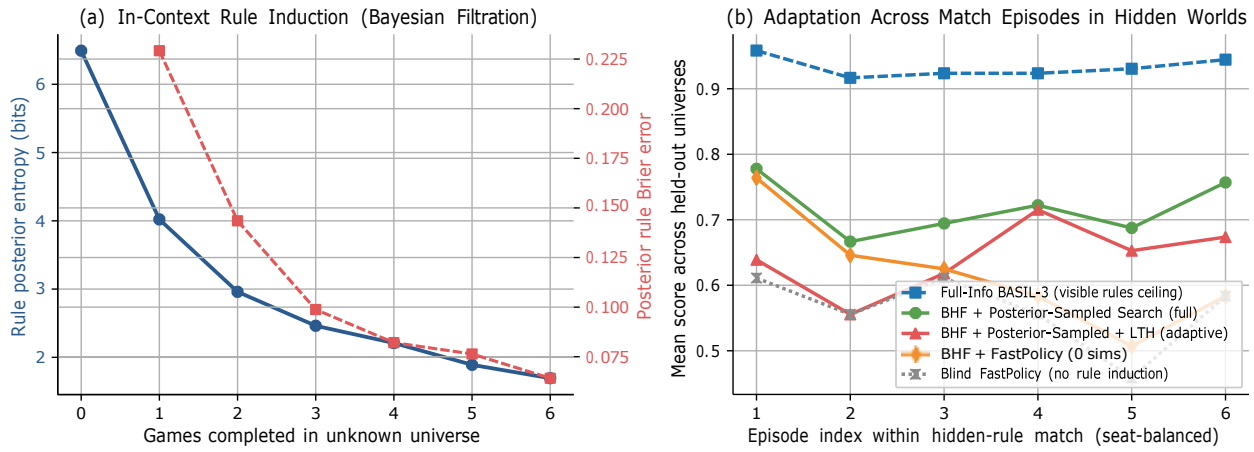


Figure 4: Hidden-rule adaptation. Left: posterior entropy and marginal rule Brier error across six games in the same universe. Right: seat-balanced score by episode for blind play, BHF with the fast policy, posterior-sampled search with and without LTH, and the visible-rule reference. The reference is an empirical comparator rather than a formal upper bound.

Table IV: Hidden-rule match results. Episode-group scores (Ep.), overall match score (Match), and mean model expansions per move. These aggregate results do not include confidence intervals.

Agent	Ep. 1–2	Ep. 5–6	Match	Exp./move
Blind FastPolicy	0.555	0.583	0.565	0.0
BHF + FastPolicy	0.646	0.583	0.604	0.0
BHF + posterior-sampled search + LTH	0.556	0.674	0.648	19.6
BHF + posterior-sampled search (full)	0.667	0.757	0.715	38.3
Visible-rule BASIL-3 reference	0.917	0.945	0.928	30.8

APPENDIX: EVALUATION IDENTIFIERS AND REPORTING CONVENTIONS

The evaluation interface uses the shorthand identifiers `morty` for FastPolicy, `citadel` for BASIL-3, `meeseeks` for LTH, and `cfc` for BHF. They identify implementation components, not additional agents in the tables. "Pre-Regret" denotes a gate evaluated before search, while "LTH-Regret" can stop at intermediate checkpoints. "VisitShare" and "UnreachStop" denote the visit-share and unreachable-leader heuristics. Scores are rounded to three decimal places, and paired improvements are calculated before rounding. Expansions count newly expanded model nodes rather than every traversal of an existing tree edge.

REFERENCES

[1] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, *et al.*, "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play," *Science*, vol. 362, pp. 1140–1144, 2018, doi: 10.1126/science.aar6404.

[2] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, *et al.*, "Mastering Atari, Go, chess and shogi by planning with a learned model," *Nature*, vol. 588, pp. 604–609, 2020, doi: 10.1038/s41586-020-03051-4.

[3] W. Ye, S. Liu, T. Kurutach, P. Abbeel, and Y. Gao, "Mastering Atari games with limited data," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 25476–25488.

[4] S. Russell and E. Wefald, *Do the Right Thing: Studies in Limited Rationality*. Cambridge, MA, USA: MIT Press, 1991.

[5] N. Hay, S. Russell, D. Tolpin, and S. E. Shimony, "Selecting computations: Theory and applications," in *Proc. 28th Conf. Uncertainty Artif. Intell.*, 2012, pp. 346–355.

[6] T. Anthony, Z. Tian, and D. Barber, "Thinking fast and slow with deep learning and tree search," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[7] J. B. Hamrick, A. J. Ballard, R. Pascanu, O. Vinyals, N. Heess, and P. W. Battaglia, "Metacontrol for adaptive imagination-based optimization," in *Proc. Int. Conf. Learn. Represent.*, 2017.

[8] I. Osband, J. Aslanides, and A. Cassirer, "Randomized prior functions for deep reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.